Examples on Variable Selection in PCA in Sensory Descriptive and Consumer Data

Per Lea, Frank Westad, Margrethe Hersleth
MATFORSK, Ås, Norway

Harald Martens
KVL, Copenhagen, Denmark

6th Sensometrics Meeting
Dortmund, July 31 - August 2, 2002
Outline of presentation

- Introduction
- Theory: Methods
- Practice: Applications
- Summary
Background

- Increasing number of measurements/data sources
- Not enough professional data analysts in the world
  \[\Rightarrow\] YOU have to analyse your own data

- Choice of method(s)
- Safe use of the methods
- Interpret - draw conclusions
- How to present results to colleagues, client, boss….
Multivariate modelling - Important aspects

- Outlier detection and their influence on the model
- Validation and model dimensionality
- Interpretation of model parameters and underlying structures
- Variable selection

⇒ Estimation of uncertainty is vital in all these matters!

“A number without any associated uncertainty is close to a random number”

- Peter Wentzell, Halifax, Canada
Bilinear models

- **One block of data ("X")**
  - Assume a model which is linear in scores and loadings; extracted in terms of *factors* (so-called latent variables)
  - The scores are linear combinations of the original variables
  - Example: Principal Component Analysis (PCA)

- **Two blocks of data ("X and Y")**
  - Regression methods which decompose the matrices in terms of factors/components
  - Examples:
    - Principal Component Regression (PCR)
    - Partial Least squares Regression (PLSR)
Validation

- **Data-model based**
  - Cross-validation (one set of objects)
    - We can validate by taking “one product out”, “one day out”, “one judge out”, “one consumer category out” etc.
  - Test set validation (two or more set of objects)

- **System/process based**
  - Validate on country level
  - Between different panels
  - … and more
Residual variance - validation

- Validation - “conservative model-fitting”
- Calibration - “numerical model-fitting”
Validation is essential

Consumer questionnaire attitudes (103×11)
Residual variance

Random numbers (103×11)
Residual variance

![Graph showing residual variance for PC and PC variance](image1.png)

![Graph showing residual variance for PC and PC variance](image2.png)
Validation is essential

Consumer questionnaire
Attitudes (103×11)
Residual variance

Random numbers (103×11)
Residual variance
Rank

- Optimal number of dimensions

- What do we mean by rank
  - Numerical rank
  - Statistical rank
  - Application specific rank (using background knowledge)
How to find the correct rank in PCA

◆ Some possible approaches:

- Bartlett’s test
- SCREE plot
- Broken stick
- Keep all eigenvalues > 1 (Kaiser’s test) (Warning: do not use this one!)
- Sum of PCs explaining > 95% of the variance
- Cross validation
- Human interpretation
Significance of loadings in PCA

- PCA is often applied as an explorative tool
- Important issues:
  - The number of relevant components
  - Which variables are significant on the components
- Resampling methods such as jack-knifing and bootstrapping are valuable tools for estimation of uncertainties in multivariate models
- Some other approaches:
  - Keep loadings > 0.3
  - Keep loadings > specified value based on number of samples (from tables based on simulations)
  - Keep subset of variables to preserve the overall information
Uncertainty estimates

◆ Objectives
  ● To estimate uncertainties in the model parameters
  ● Reflect the *actual* data structure (outliers, skewness)

◆ Some approaches for estimation (Efron and Tibshirani)
  ● Jackknifing/Cross validation (JK/CV)
  ● Bootstrapping

◆ Cross-validation for individual segments might give components that are mirrored or flipped
  \( \Rightarrow \) Restricted Procrustes rotation
Uncertainty estimates

The variance of the model parameters can be estimated by jack-knifing.

Example: Loadings, $p$

$$s^2(p) = \left( \sum_{m=1}^{M} (p - p_m)^2 \right) \left( \frac{(M - 1)}{M} \right)$$

$M$ = the number of segments
$s^2(p)$ = estimated uncertainty (variance) of $p$
$p$ = the loading using all $N$ objects
$p_m$ = the loading using all objects except the object(s) left out in cross validation segment $m$. 
Uncertainty estimates

- A univariate t-test is performed for each element $p_k$ in the loading vector relative to the square root of it’s estimated uncertainty, $s(p)$

- Use the estimates for an approximate confidence interval for each variable

- The method seems robust for various cross validation schemes (number of segments, repeated random selection)
PCA of sensory data

- Should one scale sensory data or not?
  - If not, the variables which are spanned the most will dominate
  - If scaled, small numerical differences might (erroneously) influence the result

- To reveal if scaling should be used or not, plot correlation loadings
- The correlation loadings are the correlations between the variables and the PC’s

\[
 r_{ka} = p_{ka} \sqrt{t_a^T t_a} / \sqrt{e_{0,k}^T e_{0,k}}
\]

How much is explained in PC \( a \)?

Variance before modeling starts

PCA model: \( X = TP' + E \)
Example 1: PCA on sensory descriptive data

- Product: Vanilla ice-cream
- 15 samples
- 18 sensory attributes
- Employ PCA: Three components are relevant
Scores and loadings
Vanilla Ice-cream; 15 products - 18 attributes
Correlation loadings
Ice-cream

Correlation between thickness and fattiness: 0.96

Significant on PC 1
Significant on PC 2
Significant on both
Not significant
How can we judge if the estimates are correct?

- Compare to ANOVA when “truth is known”
  - Pizza product
  - 8 samples from a $2^3$ factorial design, 29 sensory attributes
  - Analyse the data with ANOVA and PCA

- Results
  - Significant effects for 16 of the attributes (ANOVA)
  - 16 attributes significant on PC1, PC2, PC3 in the JK PCA
  - 14 of these were the same as for ANOVA
Example

- Mozzarella cheese
- 6 products for consumer test
- 105 consumers
- 3 components were found to be relevant
- Which consumers are informative? (Significance level 20%)
Correlation loadings
Mozzarella Cheese; 6 products - 105 consumers

Significant on PC 1
Significant on PC 2
Significant on both
Not significant
Summary

- Significance tests in PCA make interpretation easier

- Correlation loadings reveal the correlation structure also when variables are not scaled

- Validation is essential to assess the model dimensionality

- Restricted Procrustes is used to avoid rotation in cross-validation (flipping, mirroring)